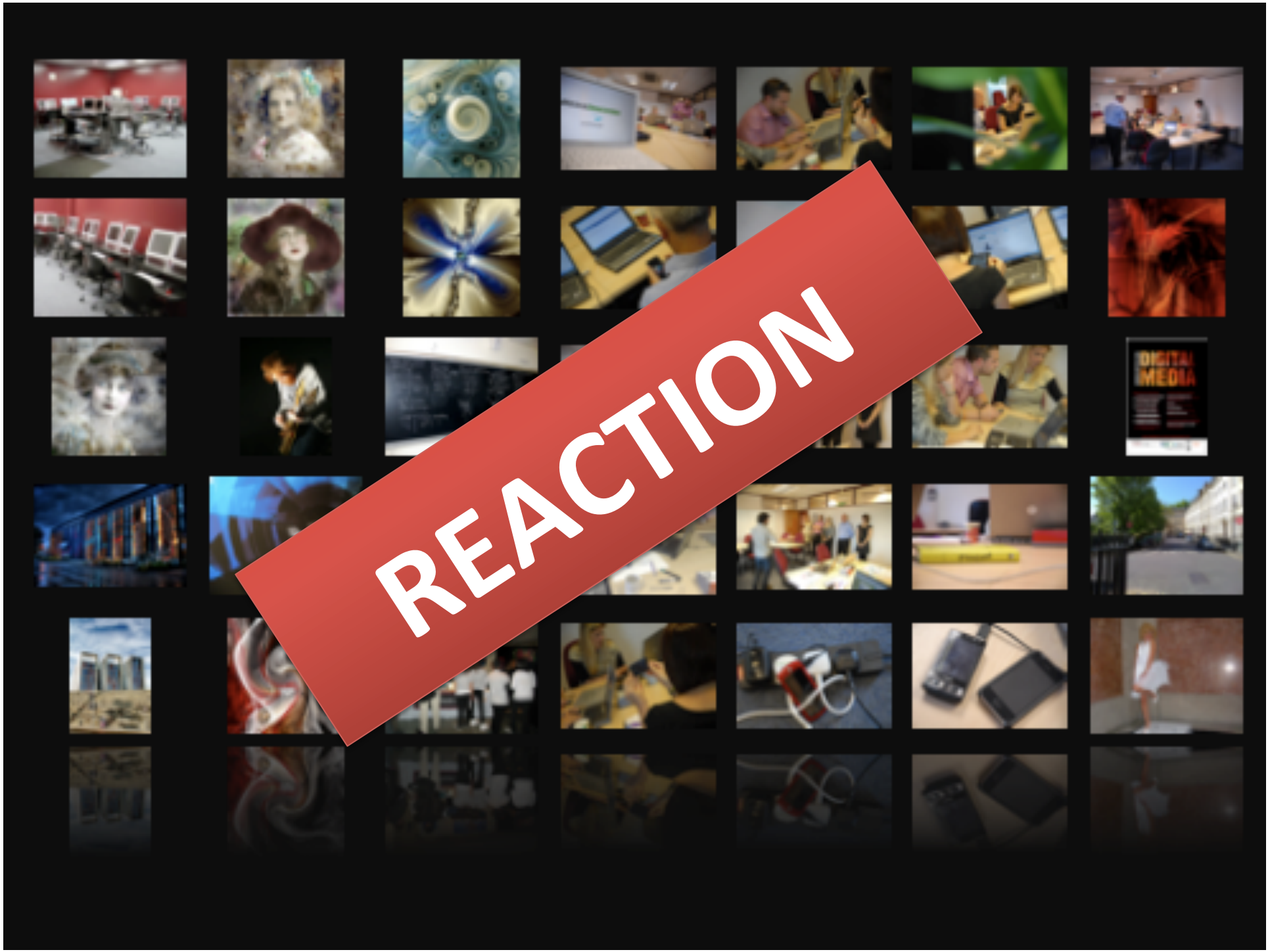
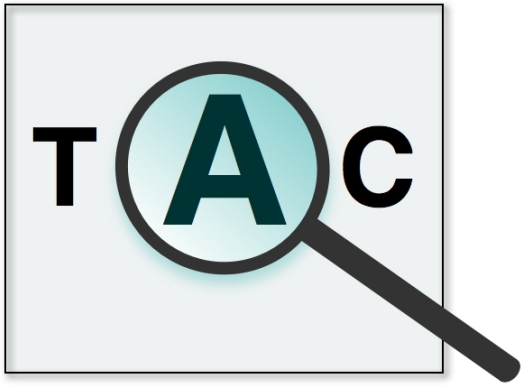

REACTION

Information Discovery

Relationship extraction techniques to support **information discovery in journalists' activities**

- **Entity Ranking**: finding the relevant entities for a given topic
- **Entity Distillation**: finding relevant resources for a given entity
- **Attribute Selection**: finding a list of key aspects to compare and differentiate a given set of entities

REACTION



TAC 2011 Knowledge Base Population (KBP2011) Track

- Entity Linking
  - +/- Entity Distillation
  - Mono-lingual, not Cross Lingual
  - Evaluation period: **8 to 14 of August**
- Slot Filling
  - Attribute Selection
  - Not enough resources to participate this year
- Temporal Slot Filling
  - Out of our goals (for now)

REACTION

Entity Linking

- Input
  - Entity name
  - Background document
- Output
  - A Knowldege Base entry (wikipedia entry)
    - to which the entity name refers
  - Or NIL if not referred in the KB

REACTION

Example

- Input

  <query id="EL000304">

  <name>Barnhill</name>

  <docid>eng-NG-31-100578-11879229</docid>

  </query>

- Output
  - NIL or KB Id

REACTION

Knowledge Base

- October 2008 dump of English Wikipedia
- Each entry has
  - a name string
  - an assigned entity type of
    - PER, ORG, GPE, or UKN (unknown)
  - a KB node ID (a unique identifier, like “E101”)
    - There are 818,741 nodes
  - a set of ‘raw’ (Wikipedia) slot names and values
  - some disambiguating text
    - (i.e., text from the Wikipedia page)

REACTION

KB Genres

Genre	Approximate Size (documents)
Broadcast Conversation	17
Broadcast News	665
Conversational Telephone Speech	1
Newswire	1,286,609
Web Text	490,596

REACTION

Training Data

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Training	2009 Eval	627	2710	567
	2010 Training Web data	500	500	500
	2010 Eval Newswire	500	500	500
	2010 Eval Web data	250	250	250
Evaluation	Newswire	500	500	500
	Web data	250	250	250

REACTION

External Resources

- Every participant
  - at least on closed submission
  - one which does not access the Web during the evaluation period
- Allowed:
  - Using a Wikipedia derived resource to (manually or automatically) create training data
  - Compiling lists of name variation based on hyperlinks and redirects before evaluation
  - Using a Wikipedia derived resource before evaluation to create a KB of world knowledge which can be used to check the correctness of facts
  - Preprocess/annotate a large text corpus before the evaluation to check the correctness of facts or aliases

REACTION

Our approach

- Pre-processing
  - POWER-TAC
    - Every KB entry representing a POWER instance
    - Every Training Data entry or entity representing a POWER Instance
  - Lucene-TAC
    - Index each KB entry
    - Index each Training Data entry
    - optimize BM25 cost function using training data

REACTION

POWER Matcher v0

- Match the input name string with all the instances in Power
  - Jaccard index on the names
- If the most similar instance is a POWER Instance
  - Representing a KB entry return its identifier
  - Otherwise return NIL
- Not using any text

REACTION

Lucene Matcher v0

- Query Lucene with the input name string
- Filter the documents whose POWER Instance was not selected by the POWER Matcher
  - Threshold on Jaccard index
- If the top document represents a POWER Instance
  - Of a KB entry return its identifier
  - Otherwise return NIL

REACTION

POWER Matcher v1

- Use synonyms
  - based on KB hyperlinks and redirects
- Use acronyms
  - Based on KB naïve parsing
    - If we have "X… Y… Z… (XYZ)"
    - Then XYZ is an acronym of X… Y… Z…
- Include other type of semantic information
  - Infobox slot-names

REACTION

Lucene Matcher v1

- Exploit the background document
  - Recognize the input name string in the background document
  - Words near the match (neighbor words) can be used for query expansion
  - Filter the neighbor words by their relevance
    - Keywords on the KB, training data and POWER

REACTION

Lucene Matcher v2

- Machine Learning Filter Approach
  - Classify if a given Lucene output entry is really related to the input name string
  - Features are
    - the neighbor words from the background document
    - Lucene output entry keywords
      - Infobox slot-names
      - words in the first paragraph of the Lucene output entry
- Train the model in the training set provided

REACTION

Work Plan

- Pre-processing
- POWER MATCHER v0
- Lucene MATCHER v0
- Submission generation and evaluate
- POWER MATCHER v1
- Lucene MATCHER v2
- Submission generation and evaluate
- Lucene MATCHER v3
- Submission generation and evaluate

REACTION